

IS

Lectures (1 to 6)

IS

* Skills of data scientist

- 1) Statistics skills.
- 2) Database.
- 3) Critical thinking, creative, Adaptive, Communication skills.
- 4) Machine learning + Data mining + Advanced Mathematics.
- 5) Collect data from different online source.
- 6) Extract data & Analysis
- 7) Programming skills.
- 8) Web (development + design).
- 9) Can make correlations & connections.

* Data enables Professional (data collectors)

↳ Quantitative (can measure upcoming data & give technical reports on it)

↳ Skeptical (be

↳ Communications & collaborative.

Big data

↳ refers to the exponential growth and availability of data, both structured and unstructured.

* Three V's describe definition of big data?

- 1) Volume 2) velocity 3) variety

1) Volume

↳ There is a large increase of data volume (why)?

- a. all of transactional data that has been added up over the years.
- b. streaming data from social media.
- c. machine to machine data increase.

← التقديرات الكبيرة ~~التي~~ تتحرك بسرعة مع
أو (data) السريعة.

2) velocity

↳ Data is being streamed at huge speeds and need to be dealt with any timely manners like (social media & mobile devices)

3) Variety

↳ many different of data

- a. Email
- b. Numeric data
- c. structured data
- d. unstructured documents
- e. Audio & video
- f. Application data.

← منظمات كثيرة جداً يتعامل على أنها تتغير
في أنواع أو (data) المختلفة.

Veracity 4 V's انهم 4 V's

* Veracity (uncertainty of data)

↳ refers to the trustworthiness of the data.
With many forms of big data (quality & accuracy)
are less controllable.

5th V's ⇒ Value of data is added

↳ well and good for access or useless data

Big data

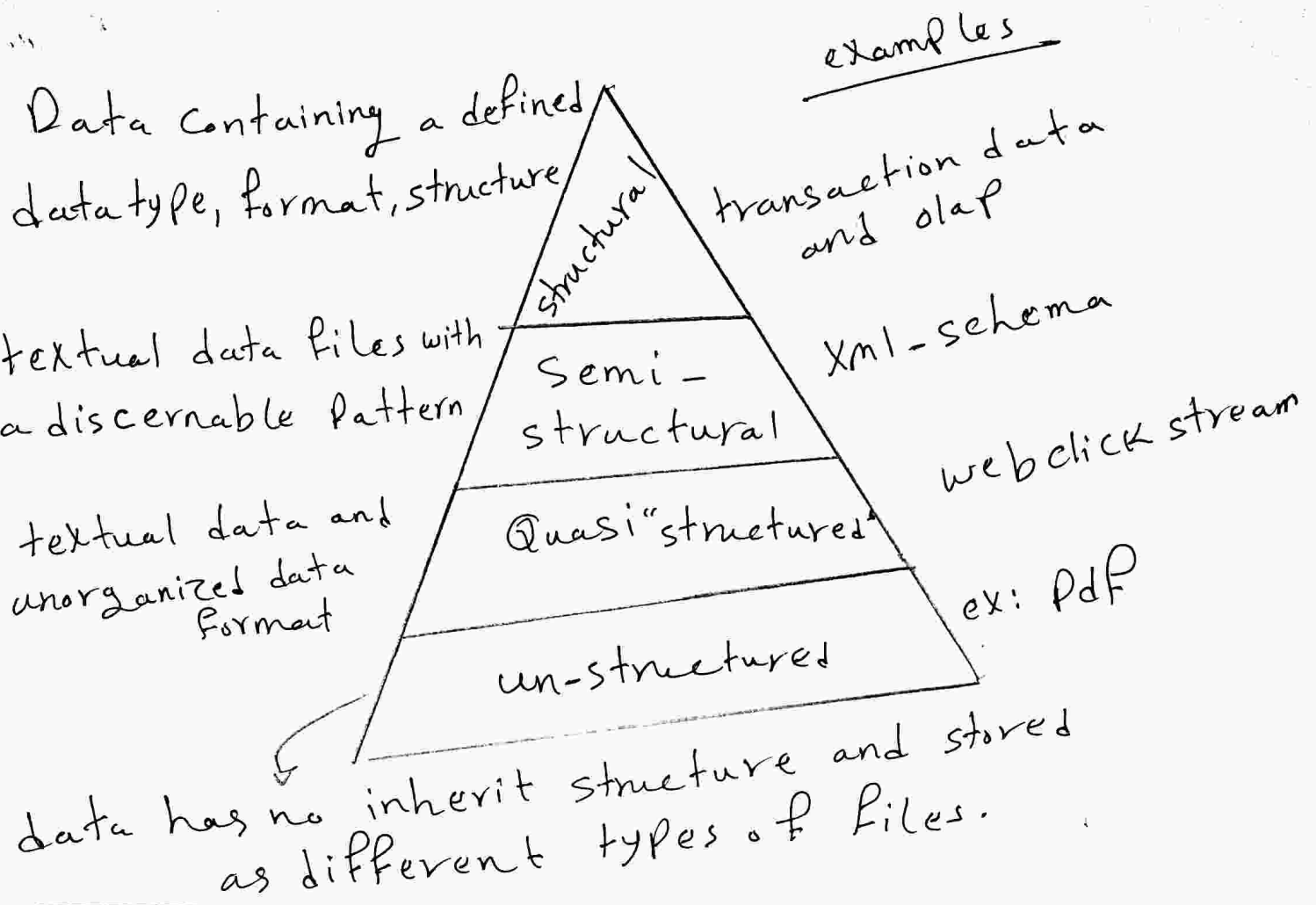
↳ data whose scale, distribution, diversity and for timeless require the use of technical architectures and analytics to enable.

* Key characteristics of Big data.

1) data volume.

2) Processing complexity

3) Data structured.



| Data islands spread sheet marks | Data warehouse | Analytic Sand box |
|---|---|--|
| is isolated data. | Centralized data containers in a purpose built space | Data assets gathered from multiple sources & technologies for analysis |
| Analyst dependent on data extracts. | Analyst dependent on IT & DBAs for data access and schema changes | analyst owned gives high performance reduce cost associated. |

| Business intelligence | Data science |
|---|---|
| → structured data, traditional sources, manageable datasets. | → structured / unstructured data & multiple types of sources & very large data sets |
| standard | optimization, predictive modeling, statistical analysis. |
| his questions did How many we we sell? Where is the problem? | What if ...? open ended questions? |

* criteria of Big Data Projects

- 1) Speed of decision making.
- 2) Throughput.
- 3) Analysis Flexibility.

* Data scientist Key Activities

- 1) reframe business^{ss} challenges as analytical challenges.
- 2) Design & implement and deploy statistical models and data mining techniques of big data.
- 3) create insights that lead to actionable recommendations.

* Three Key roles of the new data Ecosystem:-

- 1) Deep analytical talent
↳ Data science
People with advanced training in quantitative disciplines such as math, statistics, machine Learning.
- 2) Data ~~savvy~~^{savvy} Professionals
People with basic knowledge of statistics and/or machine learning who can define key questions that can be answered using advanced analytics.
- 3) Technology & data enablers
↳ People providing technical expertise to support analytics projects skill sets including computer programming & DB administrator.

* Key roles for successful Analytic Projects:-

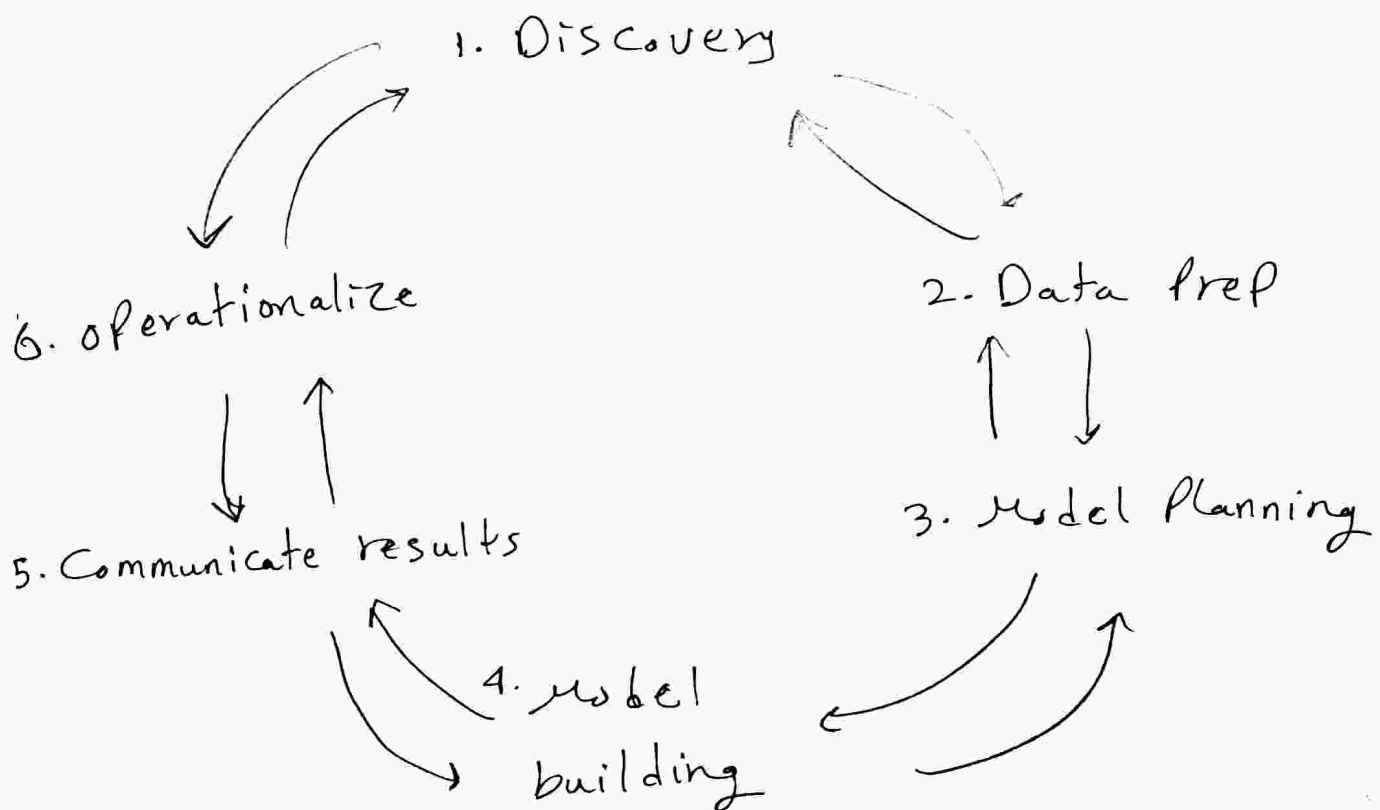
| Role | Description |
|-----------------|--|
| Business user | ↳ benefits from end data results, can consult and advise project team on value of end results. |
| Project manager | ↳ He care only about output ↳ ensure key objectives are met on time and and at expected quality. |

| | description |
|-------------------------------|---|
| Project Sponser | <ul style="list-style-type: none"> → Provide the Fund needed in Project. → Cares only about completed work. → responsible for Genesis of Projects. |
| Business intelligence Analyst | <ul style="list-style-type: none"> → measure the indicators from point of view of business |
| Data engineer | <ul style="list-style-type: none"> → responsible for data (with its variations) → extract data → has deep technical skills for data management. |
| Data administrator (DBA) | <ul style="list-style-type: none"> → Provisions and Configures database environment to support analytical needs of working team. |
| Data scientist | <ul style="list-style-type: none"> → deal with data analytically. → ensure that overall analytically objectives are met. |

* value of using data analytics lifecycle?

- 1) ensure rigidity and completeness.
- 2) enable better transition to members of the cross-function analytic teams.

Data analytics life cycle



من لو وصلت لمرحلة ولقيت ليا دل (data) ليا
معاك غير كافيه فترجع للخطة السابقة وتعيد
حساباتك.

Data Analytics lifecycle

له شغلان reverse & Forward

① Discovery

له يعرف بعد اذل خطوة هل عندي (data) كافيه
اعمل بيه (analytic plan) لو فيه حاجات ناقصة
يبقى الخطوة لسة مخلصه.

② Data Prep. (Collect data)

له هل عندي (بعد ما) (data) كافيه اعلم بيه (build for model)

③ Model Planning

له بعد ما حل اذل (planning) حوله على اساس مبرمج
له لو تمام هكمل

④ Model build

له اعلم (build) وبعدين هعمل (test for model)

⑤ Communicate results

له لازم يكون عندي (skills) تقنع بيه غيرك بالنتائج
اللى حقتها.

⑥ operationalize

1] Discovery

10

↳ Problem definition phase.

↳ need to learn about domain we are working on

↳ Know more about history of this domain

له بشوف اللي قبل عليه اديه و دكتورا لفين و ايه المشاكل

للي واجهتهم و هل التجربة نجحت ولا لا.

له هل صر علي (Project) شوية بس ..

له لازم اقيس ال (resources) اللي مالحا .

2] Data Preparation

له بعد (Prepare) ال (data)

له ال (Phase) اللى بيعد فيها (Sandbox)
 ← كذا مدرستين

extract-load-transform ← ELT (a)

extract-transform-load ← ETL (b)

~~extract data from data warehouse~~

← كذا

↳ load data in sandbox.

↳ transform (لونسك (data في (scales) مختلفة

له كذا صيغ الشغل بتاعك

له بعد (transform) في (range) واحد.

11

~~Conversion~~ ~~عمل~~ ~~(transform)~~ ~~عمل~~ ~~السرعة~~ ~~ال~~

ال (Preparation, Discovery) أكثر (2 Phases)

بمعدل فهم (refine) & ~~عمل~~

3 Model Planning

ل بناء على اللى سمعت في 1، 2 بعد اصدار البرنامج
بمعدل فهم (refine) & ~~عمل~~

ل عمل (Feature selection)

ل بفكر واحد ال (work-flow) criteria اللى
هش على.

4 Model build

ل ال (implementation)

ل عمل (test) ال (model) بمعدل

5 Communicate results

ل يعرف النتائج بمعدل ال (benefits) اللى
وصلت لهما.

6 Operationalize

ل عمل (operation) للشغل بمعدل.

* 4 Core deliverables to meet most stakeholders needs:-

1] Presentation For Project Sponsors:-

- Big picture takeaways For executive level stakeholders
- determine Key messages to aid their decision-making process.
- Focus on clean, easy visuals For Presenter to explain & for ~~the~~ viewer to grasp.

2] Presentation For analysts:-

- Business Process changes.
- reporting changes.
- ~~For~~ data scientists want the details.

3] Code: For technical People

4] Technical Specs: of implementing the code.

* Analyst wishlist For a successful analytics Project:-

1] Data & workspaces

- a. access all data.
- b. sandbox
- c. Ability to move data back between staging.
- d. up-to-date data dictionary.

[2] Tools

- a. statistical, mathematical, visual SW.
- b. tool or place to log errors with systems.
- c. Collaboration → online platform for communication between team members.

Sandbox: Data assets gathered from multiple sources and technologies for analysis.

↳ high performance analytics.

↳ reduce costs of data replication.

↳ Analyst owned.

Tools used in lifecycle

[1] Data Preparation

- descriptive statistics.

- visualization (R), Spotfire

→ for data transformation

↳ SQL, Hadoop, Mapreduce.

2] Model Planning

- R / Postgres SQL, sql analytics, Apline miner, SPSS / ODBC.

3] Model Building

↳ R, PL/R, SQL, ~~SAS~~

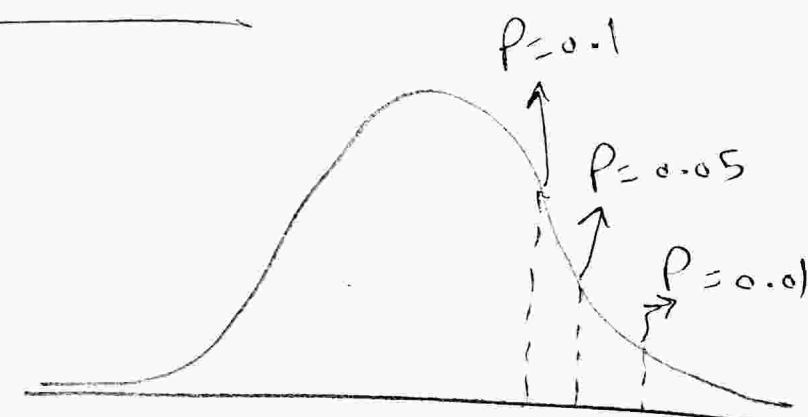
* Distribution of sample means

→ Calculate the mean

$$\frac{\sum (x - \bar{X})}{N}$$

→ calculate variance & standard deviation

$$s^2 = \frac{\sum (x - \bar{X})^2}{n}$$



→ calculate the p value

if p-value is between 0.01 & 0.1
↳ inside range (normal case)

if not → reject ~~the~~ null hypothesis.

Notes

*Significance

↳ Probability of False Positive (α)

*Power

↳ Probability of a true Positive ($1-\beta$)

*Effect size

↳ size of observed difference.